Towards a spatio-temporal deep learning approach to predict malaria outbreaks using earth observation measurements in South Asia

Usman Nazir[1], **Ahzam Ejaz**[1], **M. Talha Quddoos**[1], **Momin Uppal**[1], **Sara Khalid**[2]
[1] Lahore University of Management Sciences
{usman.nazir, ahzam.ejaz, muhammad.quddoos, momin.uppal}@lums.edu.pk
[2] University of Oxford
sara.khalid@ndorms.ox.ac.uk

Abstract

Environmental indicators can play a crucial role in forecasting infectious disease outbreaks, holding promise for community-level interventions. Yet, significant gaps exist in the literature regarding the influence of changes in environmental conditions on disease spread over time and across different regions and climates making it challenging to obtain reliable forecasts. This paper aims to propose an approach to predict malaria incidence over time and space by employing a multi-dimensional long short-term memory model (M-LSTM) to simultaneously analyse environmental indicators such as vegetation, temperature, night-time lights, urban/rural settings, and precipitation. We developed and validated a spatio-temporal data fusion approach to predict district-level malaria incidence rates for the year 2017 using spatio-temporal data from 2000 to 2016 across three South Asian countries: Pakistan, India, and Bangladesh. In terms of predictive performance the proposed M-LSTM model results in lower country-specific error rates compared to existing spatio-temporal deep learning models. The data and code have been made publicly available at the study GitHub repository: `https://github.com/usmanweb/CITY-at-LUMS`.

1 Climate Impact Statement

Climate change has been linked with acceleration and aggravation of the spread of infectious diseases; consequently worsening the risk to human health. Pathogens responsible for diseases such as malaria, cholera, Zika virus and others are known to spread more easily amidst extreme environmental changes such as heatwaves, floods, droughts, and wildfires caused by globally rising temperatures. Moreover, the spread of infectious diseases varies geographically, and this variability is connected to changes in the local environment. In the context of the South Asia use-case presented in this study, malaria incidence was predicted for each district in India, Pakistan, and Bangladesh for the year 2017 using geo-spatio-temporal data from 2000 to 2016. The proposed approach integrates environmental factors and historical malaria data, offering the ability to make localised temporal predictions. This fusion of spatial and historical information, generalisable to other infectious diseases affected by the environment, has the potential to contribute significantly to country and regional level malaria early warning systems, providing valuable insights for proactive and targeted interventions against a rapidly changing climate.

2 Introduction

Malaria remains one of the leading communicable causes of death Gelband et al. (2020); Cowman et al. (2016). Approximately half of the world's population is considered at risk, predominantly in

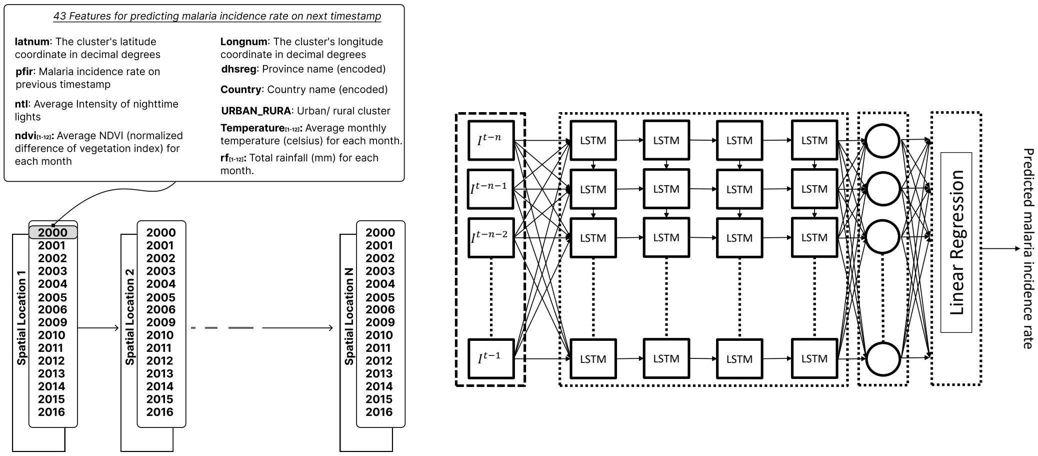
37th Conference on Neural Information Processing Systems (NeurIPS 2023), Louisiana, United States.

African and South Asian countries Rosenthal et al. (2019). Although malaria is preventable, spatio-temporal heterogeneity in climatological, sociodemographic, and environmental risk factors make outbreak prediction challenging. Data-driven approaches accounting for spatio-temporal variability may offer potential for region-specific malaria predicting tools.

Production of accurate models is complicated on two fronts. Firstly, tracking malaria incidence can present challenges, particularly in resource-limited settings. Often, these measurements are extrapolated from household surveys that are small in scale and expensive, and therefore prone to mis-estimation. Secondly, while environmental factors have been found to play a role in influencing the life cycle of the malarial parasite Gaudart et al. (2009), the fluctuation of outbreak patterns across different regions and time periods, and their dependence on multi-dimensional factors, make modelling these impacts complex. Various methodologies have been developed to address this challenge in a data-driven manner. Machine learning and deep learning techniques have gained prominence in recent years for their ability to analyze complex spatio-temporal data. For instance, the use of classification techniques such as KNN, Naive Bayes, and Extreme Gradient Boost have proven effective in examining the relationship between historical meteorological data and records of malarial cases Kalipe et al. (2018). Additionally, the integration of spatio-temporal earth observation measurements, e.g. sea surface temperature variability, has shown promise in machine learning approaches for predicting malaria outbreaks Martineau et al. (2022). We propose a deep learning method (M-LSTM) to analyze historical environmental data with the aim to provide insight into how the combination of such factors can contribute to the development of early warning systems.

2.1 Contributions

We derived a bespoke dataset on environmental indicators such as nighttime lights directly from satellite imagery. We then linked this bespoke dataset with publicly available environmental indicators datasets and malaria incidence datasets. Subsequently, we implemented a meticulous data segmentation process, dividing the data based on both spatial locations and time, providing precise and localized insights. This enabled us to leverage a comprehensive set of 43 features for each spatial location for each year included in the study. Together, these contributions resulted in improved predictive performance as well as a detailed dataset for the use of the wider research community.


43 Features for predicting malaria incidence rate on next timestamp
latnum: The cluster's latitude coordinate in decimal degrees
pfir: Malaria incidence rate on previous timestamp
ntl: Average Intensity of nighttime lights
ndvi[1-12]: Average NDVI (normalized difference of vegetation index) for each month
Longnum: The cluster's longitude coordinate in decimal degrees
dhsreg: Province name (encoded)
Country: Country name (encoded)
URBAN_RURA: Urban/ rural cluster
Temperature[1-12]: Average monthly temperature (celsius) for each month.
rf[1-12]: Total rainfall (mm) for each month.
Spatial Location 1
Spatial Location 2
Spatial Location N
2000
2001
2002
2003
2004
2005
2006
2009
2010
2011
2012
2013
2014
2015
2016
I^{t-n}
I^{t-n-1}
I^{t-n-2}
I^{t-1}
LSTM
Linear Regression
Predicted malaria incidence rate

(a) (b)

Figure 1: (a): Over the course of time, data is collected at a specific spatial point, with each time step $(2000, 2001, \ldots, 2016)$ encompassing 43 unique features at that particular spatial location. (b): The dataset begins in the year 2000. The dataframe for I^{t-n} corresponds to that year, I^{t-n-1} aligns with 2001, and successive instances continue in the same manner. Ultimately, I^{t-1} encompasses data from the final timestamp in 2016. The LSTM model comprises four hidden layers, each containing 100 LSTM units. Subsequently, a fully connected layer is employed, followed by linear regression to estimate the malaria incidence rate for the year 2017 through spatio-temporal splitting and for a random year through spatial splitting.

3 Methodology

3.1 Data Sources

1. **Malaria Incidence Rates**: We obtained data on malaria incidence rates for South Asia covering the period from 2000 to 2017. There are 1,517 spatial locations for Pakistan, 28,395 for India, and 2,260 for Bangladesh. These data were sourced from the Demographic and Health Survey (DHS) datasets provided by the US Agency for International Development.

2. **Environmental Factors**: We collected data on environmental factors linked to malaria outbreaks, including temperature (measured in degrees Celsius), rainfall (measured in millimeters), and the average Normalized Vegetation Index for each month from year 2000 to 2017. These environmental data were extracted from the Advancing Research on Nutrition and Agriculture (AReNA) project conducted by the International Food Policy Research Institute (IFPRI) in 2020, as documented in `https://www.ifpri.org/`.

3. **Nighttime Lights Data**: Additionally, data pertaining to nighttime lights were gathered from two distinct satellites. The first satellite, known as DMSP OLS (Defense Meteorological Program Operational Linescan System), provided nighttime lights data for the years 1992 to 2014. The second satellite, VIIRS (Visible Infrared Imaging Radiometer Suite) Nighttime Day/Night Band Composites Version 1, offered data spanning from 2012 to 2023.

3.2 Modeling pipeline

The architecture for forecasting malaria incidence rates comprised a structured approach using a four-layer M-LSTM model, comprising 100 LSTM units per layer (see ablative study in Appendix A), alongside a fully connected regression layer (see Fig. 1: (b)). The modelling process included the collection of time-series environmental data and historical malaria rates. Feature engineering created a dataset with temporal sequences of environmental data and corresponding malaria rates (see Fig. 1: (a)). Data division into training, validation, and test sets ensured robust model evaluation. The core methodology involved constructing a four-layer LSTM network to capture temporal patterns, with training focusing on minimizing prediction errors. Hyperparameter tuning optimized model parameters, while validation assessed generalization. Testing yielded objective performance metrics.

3.3 Data Pre-processing

Data preprocessing involved the following steps:

1. **Data Cleaning and Merging**: We conducted a comprehensive data cleaning and merging process to consolidate information from both nighttime lights datasets. Nighttime light intensity was quantified using digital units, ranging from 0 (indicating the absence of light) to 63 (representing the highest level of light intensity).

2. **Temporal Batching**: The data for each country was organized into distinct batches, with each batch containing 18 data points and 43 features (1 NTL, 12 NDVI, 12 Temperature, 12 Rainfall, 1 Latitude, 1 Longitude, 1 Urban/Rural, 1 Prev Malari value, 1 country, 1 Province). These 18 instances precisely corresponded to the years from 2000 to 2017. Furthermore, each batch was associated with specific geographical coordinates, including longitude and latitude.

3. **Training and Testing Split**: Out of the total 1,517 spatial locations for Pakistan, 28,395 for India, and 2,260 for Bangladesh, 80% were assigned for training and model validation, with the remaining 20% (In terms of space or time) earmarked for testing.

These preprocessing steps were essential in readying the data for subsequent analysis and modeling, guaranteeing its proper organization and suitability for the application of artificial intelligence methods (M-LSTM, Conv-LSTM, and Random Forest), as detailed in section 4.

Country	Country-specific Model	Split-type	Prediction-year	R-squared (R^2)	RMSE
Pakistan	M-LSTM	Spatio-temporal	2017	0.33	0.0007
India	M-LSTM	Spatio-temporal	2017	0.99	$4.86e^{-6}$
Bangladesh	M-LSTM	Spatio-temporal	2017	0.10	$1.32e^{-5}$
Pakistan	Conv-LSTM Shi et al. (2015)	Spatio-temporal	2017	0.09	0.0128
India	Conv-LSTM Shi et al. (2015)	Spatio-temporal	2017	0.18	0.0162
Bangladesh	Conv-LSTM Shi et al. (2015)	Spatio-temporal	2017	0.01	0.0016
Pakistan	M-LSTM	Spatial	Random	-	0.0076
India	M-LSTM	Spatial	Random	-	0.0166
Bangladesh	M-LSTM	Spatial	Random	-	0.0016
Pakistan	Random Forest Breiman (2001)	Spatial	Random	-	0.026
India	Random Forest Breiman (2001)	Spatial	Random	-	$5.15e^{-5}$
Bangladesh	Random Forest Breiman (2001)	Spatial	Random	-	0.0009

Table 1: Experimental Findings: The R-squared and Root Mean Square Error (RMSE) of M-LSTM, Conv-LSTM, and Random Forest models are assessed on a testing country-specific datasets with a space and time-based split.

4 Quantitative Evaluation

4.1 Performance Metrics

Model fit was assessed using R-squared (R^2) and RMSE. For a given district-level location and year, let X denote true malaria incidence rate, and let $\hat{X}$ denote the corresponding model-estimated malaria incidence rate.

$$RMSE = \sqrt{\frac{\sum_i^N (X_i - \hat{X}_i)^2}{N}} \,, \tag{1}$$

$$R^2 = \frac{1 - \sum_i^N (X_i - \hat{X}_i)^2}{\sum_i^N (X_i - \hat{X}_i)^2} \,, \tag{2}$$

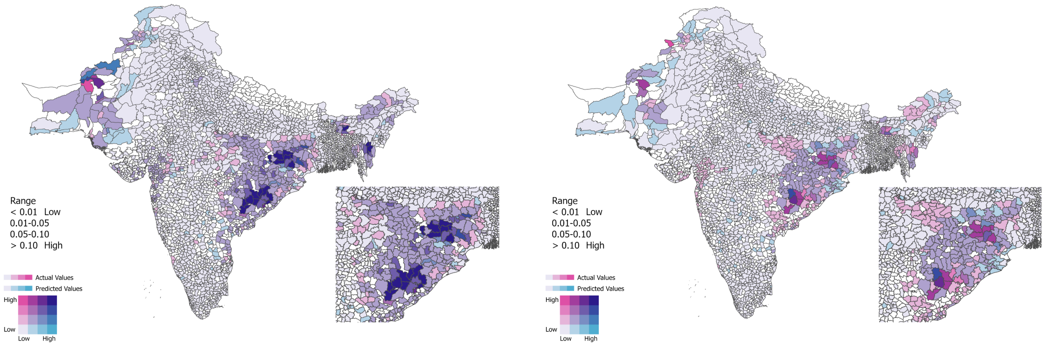
We developed and verified the Multidimensional LSTM model, contrasting its performance with an existing spatio-temporal model Conv-LSTM model Shi et al. (2015), and an existing classical regression model: Random Forest (with 100 trees) Breiman (2001). The results displayed in Table 1 demonstrate that the Multi-dimensional LSTM (M-LSTM) model surpasses the Conv-LSTM and Random Forest methods. Generally, higher R-squared and reduced error rates are attainable with increased LSTM layers and 2017 prediction year. The M-LSTM model, incorporating dense input connections, reduced RMSE (on average) by 29.41% compared to the Conv-LSTM model.

The evaluation metrics (see ablative study in Appendix A and evaluation metrics in Appendix C for details), together provided a comprehensive evaluation of the model's predictive performance. Data partitioning played a pivotal role in our evaluation strategy. The dataset underwent division into "spatial" and "spatio-temporal" splits for both training and testing. In the spatial split, an 80% training and 20% testing split was upheld for forecasting the malaria incidence rate in a random year. Conversely, in the spatio-temporal split, the training set comprised environmental data and malaria incidence rates from 2000 to 2016, with testing conducted to predict the malaria incidence rate for the year 2017. This approach ensured comprehensive evaluation across different data distributions, enabling us to assess the model's robustness and performance under varying time and space conditions.

5 Qualitative Evaluation

5.1 Bivariate Maps

In the qualitative assessment, bivariate maps comparing true and predicted malaria incidence rates are depicted in Fig.2. In the case of the proposed model, the true and predicted values exhibit a consistent correlation. For instance, in areas where the true malaria incidence rate is high, the predicted malaria incidence rate also tends to be high, as observed in regions with dark blue coloring in the magnified image of Fig. 2: (a). However, in the Conv-LSTM model, there are instances where the true values



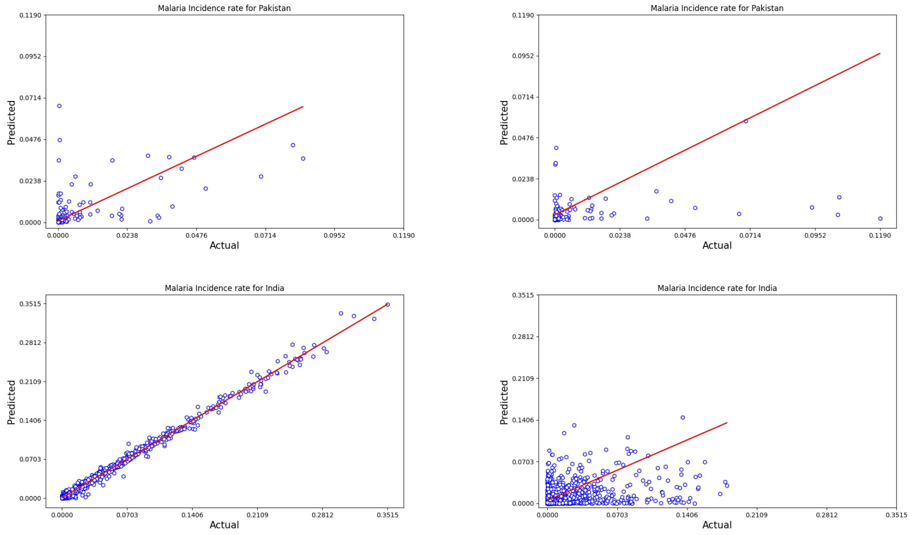
(a) Proposed M-LSTM (b) Conv-LSTM Shi et al. (2015)

Figure 2: Predicted versus true malaria incidence rates for each district of Pakistan, India, and Bangladesh. These findings are based on the analysis of the 20% country-specific test dataset.

for certain regions are high, but the predicted values are low, as indicated by the dark pink coloring in the zoomed image of Fig. 2: (b).

5.2 Scatter Plots

Fig. 3 illustrates the true and predicted values of malaria incidence rates for the year 2017.



Proposed Model Conv-LSTM

Figure 3: Scatter plots illustrating actual and predicted malaria incidence rates, with the first row depicting Pakistan and the second row representing India.

6 Conclusion

In this paper, we introduced a South Asia case study where we predicted annual district-level malaria incidence rates using historical data from previous years. We used a data-driven approach to fuse multi-dimensional environmental and historical malaria information without recourse to prior modelling assumptions. The methods developed in this paper may empower decision-makers to study and predict malaria and ultimately other infectious disease outbreaks early in time for regional and international planning against climate-induced environmental changes in local environments.

References

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Cowman, A. F., Healer, J., Marapana, D., and Marsh, K. (2016). Malaria: biology and disease. *Cell*, 167(3):610–624.

Gaudart, J., Touré, O., Dessay, N., Dicko, A. L., Ranque, S., Forest, L., Demongeot, J., and Doumbo, O. K. (2009). Modelling malaria incidence with environmental dependency in a locality of sudanese savannah area, mali. *Malaria journal*, 8:1–12.

Gelband, H., Bogoch, I. I., Rodriguez, P. S., Ngai, M., Peer, N., Watson, L. K., and Jha, P. (2020). Is malaria an important cause of death among adults? *The American Journal of Tropical Medicine and Hygiene*, 103(1):41.

Kalipe, G., Gautham, V., and Behera, R. K. (2018). Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis. In *2018 International Conference on Information Technology (ICIT)*, pages 33–38. IEEE.

Martineau, P., Behera, S. K., Nonaka, M., Jayanthi, R., Ikeda, T., Minakawa, N., Kruger, P., and Mabunda, Q. E. (2022). Predicting malaria outbreaks from sea surface temperature variability up to 9 months ahead in limpopo, south africa, using machine learning. *Frontiers in Public Health*, 10:962377.

Rosenthal, P. J., John, C. C., and Rabinovich, N. R. (2019). Malaria: how are we doing and how can we do better? *The American journal of tropical medicine and hygiene*, 100(2):239.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28.

7 Supplementary Material

A Ablative Study

As shown in Table 2, we investigated the model performance fluctuations across different epochs, LSTM layers, and hidden units. The model we have chosen as our primary focus is prominently highlighted in red.

Epochs	LSTM layers	Hidden Units	RMSE
10	1	100	0.0127
10	2	100	0.0081
10	3	100	0.0095
10	**4**	**100**	**0.0076**
10	1	200	0.0123
10	2	200	0.0088
10	3	200	0.0074
10	4	200	0.0093
20	4	200	0.0084

Table 2: Exploratory analysis of malaria incidence rate for Pakistan for prediction on random years: Investigating variations in epochs, LSTM layers, and hidden units. Our selected model is highlighted in red.

B Earth Observation (Nighttime Lights, NDVI, etc.) and DHS Malaria Data

The Advancing Research on Nutrition and Agriculture (ARENA) project, funded by the Bill and Melinda Gates Foundation, is a six-year initiative implemented from 2015 to 2020 in South Asia and sub-Saharan Africa. It aims to close significant knowledge gaps by conducting policy-relevant

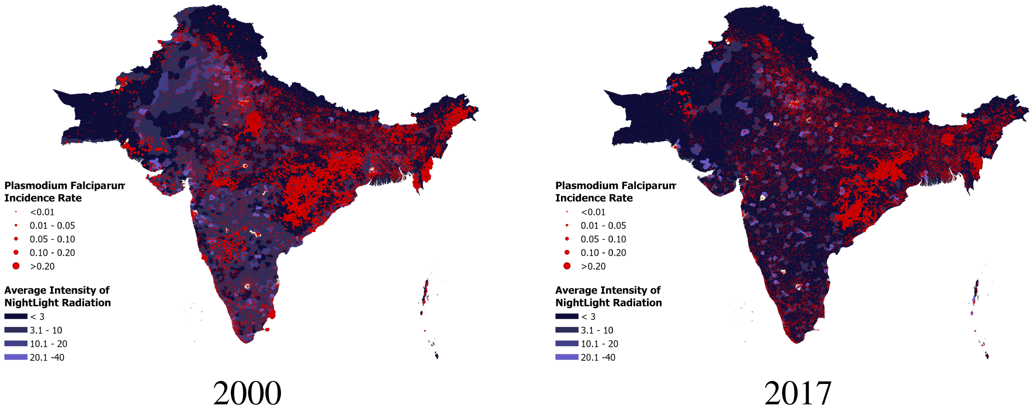


Figure 4: Nighttime lights vs DHS Malaria Incidence Rate comparison in three South Asian countries: Pakistan, India and Bangladesh. The presence of high nighttime light intensity indicates a lower likelihood of experiencing malaria disease at the district level, and conversely, areas with low nighttime light intensity are associated with a higher risk of such burdens.

research at scale and generating datasets and analytical tools to benefit the broader research community. One of the major challenges in agriculture and nutrition research is the scarcity of data. Existing datasets that include both agriculture and nutrition information are often limited in size and geographical scope. To address this, the ARENA team has constructed a large multi-level, multi-country dataset by combining individual and household-level data from the Demographic Health Surveys (DHS) with geo-referenced data on agricultural production, agroecology, climate, demography, and infrastructure.

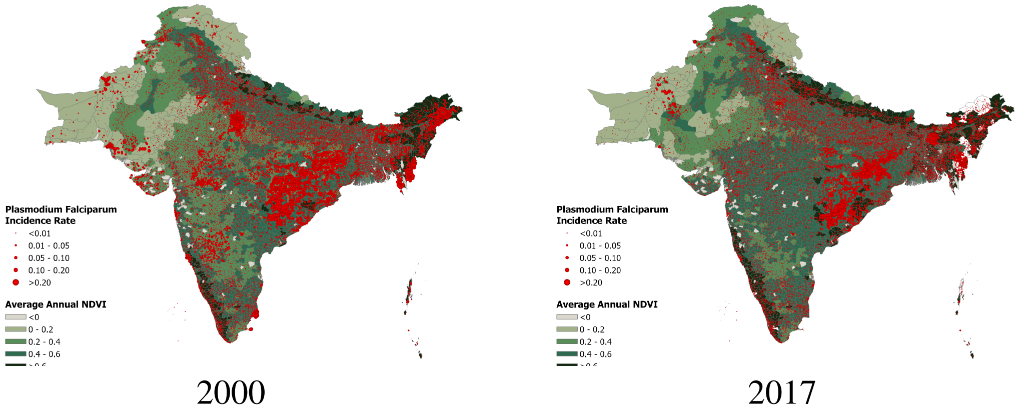


Figure 5: NDVI vs DHS malaria incidence rate comparison in three South Asian countries: Pakistan, India, and Bangladesh. Low NDVI values indicate a decreased probability of encountering malaria disease at the district level, while areas with high NDVI values are associated with an elevated risk of such burdens.

The DHS Program [1] has been collecting and disseminating accurate and representative data on population, health, HIV, and nutrition since 1984. These nationally representative household surveys have been conducted in over 90 countries through more than 400 surveys. The DHS surveys cover a wide range of health and demographic indicators, including fertility and mortality rates, family planning, maternal and child health, vaccination, prevalence of anemia, literacy rates, water and sanitation, domestic violence, women's empowerment, and tobacco use. The surveys also collect GPS location data for surveyed clusters, allowing for analysis in conjunction with additional geospatial

[1] https://www.dhsprogram.com/

variables. This section discusses the integration of two types of data sources: the DHS (Demographic

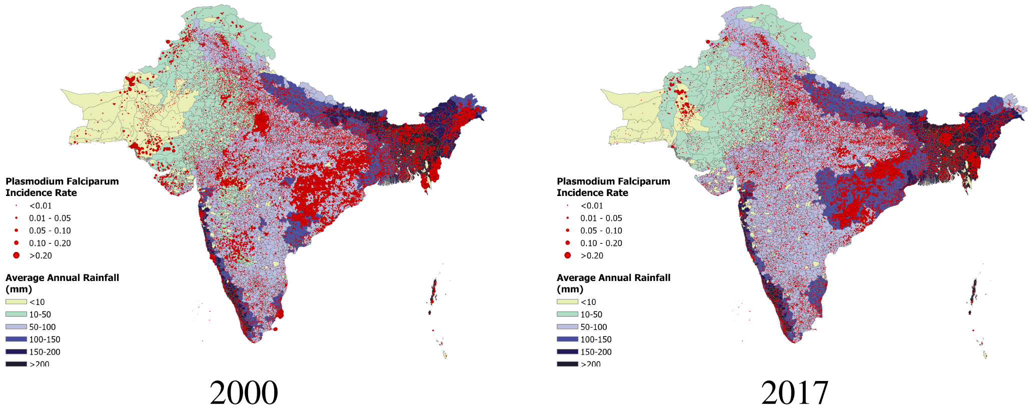


Figure 6: Rainfall vs DHS malaria incidence rate comparison in three South Asian countries: Pakistan, India, and Bangladesh. Low rainfall values suggest a reduced likelihood of experiencing malaria disease at the district level, while areas with high rainfall values are linked to an increased risk of such burdens.

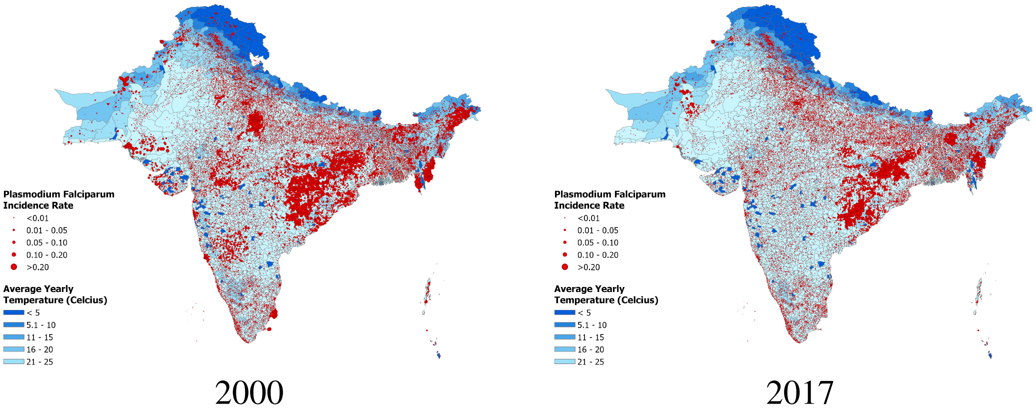


Figure 7: Temperature vs DHS malaria incidence rate comparison in three South Asian countries: Pakistan, India, and Bangladesh. Low temperatures suggest a reduced probability of encountering a malaria disease at the district level, whereas regions with high temperatures are linked to an increased risk of such burdens.

and Health Survey) surveys, which primarily provide information about health-related issues and are associated with geographic coordinates, and standard earth observation data, which includes information like nighttime lights (used as a proxy for poverty), NDVI (Normalized Difference Vegetation Index), precipitation, and temperature.

The key idea here is that these two sets of data can be combined or linked together, allowing researchers to conduct cross-country and country-level studies on various topics related to the relationship between poverty and health.

By integrating DHS survey data, which typically provides insights into health outcomes and conditions in specific geographic areas, with earth observation data, which provides environmental and socio-economic indicators such as nighttime lights (used as a proxy for poverty), NDVI (a measure of vegetation), precipitation, and temperature, researchers can gain a more comprehensive understanding of how poverty and environmental factors influence health outcomes. This integrated approach facilitates research that examines these relationships on both a broad cross-country scale and at the individual country level. It enables researchers to explore questions like how environmental conditions and poverty levels impact health disparities across different regions and countries, ultimately contributing to a deeper understanding of the poverty-health nexus.

As depicted in figures 4, 5, 6 and 7, a comparison is made between the malaria incidence rate and earth observation Indicators, including Nighttime Lights Intensity, NDVI, precipitation and temperature respectively.

C Evaluation Metrics

In-depth analysis of the model's performance encompassed several critical dimensions. Firstly, we explored the impact of varying the number of training epochs, conducting experiments with epoch counts of 10, 20, 30 and 100. Secondly, we closely monitored the behavior of the training loss throughout diverse configurations. We observed that this metric exhibited fluctuations across different training scenarios, consistently demonstrating a decrease with an increasing number of training epochs. In tandem, the model's R-squared exhibited remarkable variability, spanning a broad range. These variations in R-squared underscored the model's adaptability to different training settings. Additionally, we examined the Root Mean Square Error (RMSE), another fundamental measure of prediction accuracy. RMSE's behavior paralleled the trends observed in the other error metrics. In the majority of configurations, RMSE values consistently remained low, reinforcing the model's reliability in generating precise predictions.

Moreover, we considered the architectural aspects of the model, focusing on the LSTM units and layers. The model incorporated LSTM units with varying counts, ranging from 100 to 200, and LSTM layers, which could be configured as either 1, 2, 3 or 4. These architectural choices were systematically factored into the evaluation process, shedding light on their impact on predictive performance.